

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

## **IMAGES ARE BEST AVAILABLE COPY.**

As rescanning documents *will not* correct images,  
Please do not report the images to the  
Image Problem Mailbox.

**THIS PAGE BLANK (USPTO)**

89



Europäisches Patentamt  
European Patent Office  
Office européen des brevets



Publication number: **0 439 693 A2**

**EUROPEAN PATENT APPLICATION**

Application number: **90120874.4**  
Date of filing: **31.10.90**  
Int. Cl.<sup>5</sup>: **G06F 11/00**

Priority: **02.02.90 US 474440**  
Date of publication of application:  
**07.08.91 Bulletin 91/32**  
Designated Contracting States:  
**DE FR GB**

Applicant: **International Business Machines Corporation**  
**Old Orchard Road**  
**Armonk, N.Y. 10504(US)**  
Inventor: **Baum, Richard Irwin**  
**5 Arbor Hill Drive**  
**Poughkeepsie, New York 12603(US)**  
Inventor: **Brotman, Charles H.**  
**13 Saint Annes Road**  
**Poughkeepsie, New York 12601(US)**  
Inventor: **Rymarczyk, James Walter**  
**6 Dara Lane**  
**Poughkeepsie, New York 12601(US)**  
Representative: **Jost, Ottokarl, Dipl.-Ing.**  
**IBM Deutschland GmbH Patentwesen und**  
**Urheberrecht Schönaicher Strasse 220**  
**W-7030 Böblingen(DE)**

**Multiprocessing packet switching connection system having provision for error correction and recovery.**

A large number of processing elements (604) (e.g. 4096) are interconnected by means of a high bandwidth switch (606). Each processing element (604) includes one or more general purpose micro-processors (1202), a local memory (1210) and a DMA controller (1206) that sends and receives messages through the switch (606) without requiring processor intervention. The switch (606) that connects the processing elements is hierarchical and comprises a network of clusters. Sixtyfour processing elements (604) can be combined to form a cluster and and sixtyfour clusters can be linked by way of a Banyan network. Messages are routed through the switch (606) in the form of packets which include a command field, a sequence number, a destination address, a source address, a data field (which can include subcommands), and an error correction code. Error correction is performed at the processing elements. If a packet is routed to a non-present or non-functional processor, the switch (606)

reverses the source and destination field and returns the packet to the sender with an error flag. If the packet is misrouted to a functional processing element (604), the processing element (604) corrects the error and retransmits the packet through the switch (606) over a different path. In one embodiment, each processing element can be provided with a hardware accelerator for database functions. In this embodiment, the multiprocessor of the present invention can be employed as a coprocessor to a 370 host and used to perform database functions.

**EP 0 439 693 A2**

# MULTIPROCESSING PACKET SWITCHING CONNECTION SYSTEM HAVING PROVISION FOR ERROR CORRECTION AND RECOVERY

## BACKGROUND OF THE INVENTION

### a. FIELD OF THE INVENTION

This invention relates to the field of multiprocessing systems and error recovery in multiprocessing systems.

### b. RELATED ART

A multiprocessing system (MPS) is a computing system employing two or more connected processing units to execute programs simultaneously. Conventionally, multiprocessing systems have been classified into a number of types based on the interconnection between the processors.

A first type of conventional multiprocessing system is the "multiprocessor" or "shared memory" system (Fig. 1). In a shared memory system, number of central processing units 102-106 are interconnected by the fact that they share a common global memory 108. Although each central processing unit may have a local cache memory, cross cache validation makes the caches transparent to the user and the system appears as if it only has a single global memory.

Shared memory systems also take the form of multiple central processing units sharing multiple global memories through a connection network. An example of such a system is an Omega network (Fig. 2). In an Omega network a plurality of switches S01-S24 organized into stages route data between a plurality of processors P0-P7 and a plurality of global memories M0-M7 by using a binary destination tag generated by a requesting processor. Each stage of switches in the network decodes a respective bit of the tag to make the network self-routing. The Omega network thereby avoids the need for a central controller.

A common characteristic of shared memory systems is that access time to a piece of data in the memory is independent of the processor making the request. A significant limitation of shared memory systems is that the aggregate bandwidth of the global memory limits the number of processors that can be effectively accommodated on the system.

A second type of commonly known multiprocessing system is the multicomputer message passing network (Fig. 3). Message passing networks are configured by interconnecting a number of processing nodes. Each node 302-308 includes a central processing unit and a local memory that is not globally accessible. In order for an applica-

tion to share data among processors the programmer must explicitly code commands to move data from one node to another. In contrast to shared memory systems, the time that it takes for a processor to access data depends on its distance (in nodes) from the processor that currently has the data in its local memory.

In the message passing network configuration of Fig. 3, each node has a direct connection to every other node. Such configurations are, however, impractical for large number of processors. Solutions such as hypercube configurations have been conventionally used to limit the largest distance between processors. In any event, as the number of processors in the network increases the number of indirect connections and resulting memory access times will also tend to increase.

A third type of multiprocessing system is the hybrid machine (Fig. 4). Hybrid machines have some of the properties of shared memory systems and some of the properties of message passing networks. In the hybrid machine, a number of processors 402-406, each having a local memory, are connected by way of a connection network 408. Even though all memories are local, the operating system makes the machine look like it has a single global memory. An example of a Hybrid machine is the IBM RP3. Hybrid machines can typically provide access to remote data significantly faster than message passing networks. Even so, data layout can be critical to algorithm performance and the aggregate communications speed of the connection network is a limit to the number of processors that can be effectively accommodated.

A variant on multiprocessing system connection networks is the cluster-connected network (Fig. 5). In a cluster-connected networks, a number of clusters 502-508, each including a group of processors 510-516 and a multiplexer/controller 518, are connected through switch network 520. The cluster network has advantages over the topology of Fig. 4 in that a larger number of processors can be effectively connected to the switch network through a given number of ports. One constraint of cluster connected networks is that the bandwidths of both the cluster controller and the switch are critical to system performance. For this reason, the design of the switch and cluster controller are important factors in determining maximum system size and performance.

## SUMMARY OF THE INVENTION

It is a first object of this invention to improve

the performance of cluster-connected multiprocessing systems.

It is a second object of this invention to provide an efficient system for hard and soft error recovery in systems connected by way of a connection network.

It is a third object of this invention is to provide a computer system capable of performing complex ad hoc queries against a relational database at speeds which are several orders of magnitude faster than with today's largest mainframe computers.

In accordance with the above objectives there is provided an improved multiprocessing system and method.

In a first embodiment, an improved cluster controller is provided. The improved cluster controller includes a switch for distributing packets received from the processing elements in accordance with a destination address and packet priority, a global storage, queues for controlling packet flow to the processing elements, an assembly buffer for assembling data from the processing elements into packets, and selection logic for selecting packets from any of the assembly buffer and the global storage to the switching network.

In a second embodiment, a system and method for recovering from errors in the destination field of data being transferred between two nodes of a multiprocessing system having at least three nodes is provided. When data is misrouted to an improper node due to an errors in a destination address field, the error is detected and corrected. Once the error is corrected data is rerouted to the correct node by way of an independent data path (i.e. one other than the one on which it was received). Advantageously, this enables recovery from both soft and hard errors in the destination address field.

In a third embodiment a multiprocessor network is provided. The network is architected as a plurality of cluster controllers which connect groups of processors by way of a switch. The processing elements each include a local memory which is accessible by each of the processors in the system.

In a fourth embodiment, a packet format for use in a cluster connected multiprocessing system is provided. The packet format includes a data field, source and destination fields, a field that can cause a write into a global memory of a cluster controller, and error correct/detect fields.

#### FEATURES AND ADVANTAGES

1. The connection network design of the present system employs mainfram technology to achieve a high bandwidth system interconnection that is beyond the capabilities of many contemporary systems. High density packaging

enables the use of wide buses (e.g. 180 bits), and high speed bipolar logic allows very high frequency system clocking (e.g. 5ns). A sustained band-width of 200GB/second is achievable for uniform random message transfers.

2. A DMA Controller in each processing element provides efficient transmission of messages through a novel packet protocol, which also enables the direct addressing of non-local memories. The latter capability is important for some software algorithms that assume a shared memory structure, and is also advantageous for system debugging and service functions.

3. The interleaving of packets from multiple messages by the DMA controller effectively randomizes the pattern of packet transmissions and is important to achieving maximum bandwidth through the switch.

4. The connection network design for packet switching provides efficient message broadcasting, and global storage for control functions, in addition to basic point-to-point message transmission.

5. The packet format allows robust error handling. The use of ECC together with the source (SRC) and destination (DST) identifiers in every packet permits efficient error correction or handling. If a hardware error results in the misrouting of a packet, then one of two cases exist: (1) the packet gets misrouted to a non-existent or non-operational processing element, in which case the cluster controller reverses the SRC and DST fields and returns the packet to its sender with an error flag; or (2) the packet gets misrouted to a functional processing element, which will retransmit the packet (after applying ECC as required). Retransmission can overcome soft errors and, in case 2 above, it can also circumvent some hard failures by employing a different hardware path.

6. This highly parallel processing structure, with its high bandwidth interconnection, is well suited for a wide variety of applications, some examples of which include database processing, logic simulation, and artificial intelligence.

#### BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of a prior art shared memory system.

Fig. 2 is a block diagram of a prior art shared memory system configured using an Omega interconnection network.

Fig. 3 is a block diagram of a prior art message passing network.

Fig. 4 is a block diagram of a prior art hybrid system.

Fig. 5 is a block diagram of a prior art

switch/queues. The interconnection of a typical one of these switch/queues 802 is illustrated. Each of the input ports on each 8X8 switch/queue is bused to all eight 8X1 switch/queues. Each 8X1 switch queue can take from 0 to 8 of its inputs (quintword packets) and enter them into a single fifo output queue in each cycle of the network clock. In the same cycle a single packet (the top of queue) can be taken off the queue and passed on to the next stage of the switch network or to the final destination. If the queue is empty at the start of a cycle, a valid input packet can bypass the queue and go directly to the output, thus saving a cycle which would have been otherwise wasted in unneeded staging.

Each packet carries with it its own destination address. The addressing mechanism provides the following function. Only those packets properly addressed for the output port represented by a given switch queue will be actually be enqueued on that port. In addition, each packet will be enqueued on only one queue. The addresses must be such that an address corresponds to a unique path between a source and destination. Groups of 3 bits within each address represent the local addresses within each switch. A fixed priority scheme is used to determine in what order each of the simultaneous input packets is enqueued. Although a more sophisticated scheme could be used, since every quintword packet has the opportunity to get on the queue on every cycle, the fixed priority scheme is inherently a "fair" one (i.e., no single source will get more or less than its share of entries on the queue, unless other sources have no data for this output port.)

Fig. 9 is a more detailed diagram of the typical switch/queue 802 shown in Fig. 8. Each switch/queue contains a queue 902 of up to 64 packets. Each packet is a quintword (180 bits) in size. Each word includes 32 bits of data plus 4 bits of ECC. A packet from an input port is selected by the recognition logic 904 of a single switch/queue based on the destination address (DST id) which is contained in the control word portion of the packet. Up to eight packets (one from each input port) may be enqueued at a given output port during each cycle. Simultaneously, each output port can select a packet for transmission, either from its local queue 902, or from short circuit logic 906 which enables a single input to go directly to the output port register 910 when the queue is empty. Busy logic 908 is provided to prevent forwarding a packet when a downstream queue is full. This design prevents an output from appearing to be busy during bursts of activity, and can thereby avoid propagating the busy condition to senders.

As an example of operation, let us assume that three of the eight inputs to the 8X8 switch have

valid addresses which direct them to the second output port. The recognition logic 904 will select on those three addresses to be gated to this part of the switch. If the output port queue 902 is not empty and is not full, the the input packets will be enqueued. If the output port queue 902 is full, the Busy Logic 908 will prevent the ingating of the packets. If the output port queue 902 is empty, the Short Circuit Logic 906 will take one of the three input packets, in accord with a conventional priority scheme, and pass it directly to the output port register 910, at the same time enqueueing the remaining two packets on the output port queue. The packet in the Output Port Register 910 will be gated to the next level of the switch as long as that level is not busy.

Fig. 10 is a more detailed illustration of an exemplary one of the cluster controllers 602(1)-602(32) of Fig. 6. Cluster controller 1 602(1) will be used by way of example. Coming from the second stage of the switch network (switches 710-716), data received on the input bus 608(1) is routed to a 9 from 6 switch 1002. The 9 from 6 switch 1002 receives six inputs: one from the switching network 606, one from a global store 1004 and four from a cluster controller assembly buffer 1006. The 9 from 6 switch 1002 distributes the received data (from the six inputs) to the appropriate "octant" or to the global store 1004. The global store 1004 can be used for a variety of functions including sharing status between processing elements, process coordination, shared algorithm control, and shared data.

In order to route the received data to the appropriate octants the 9 from 6 switch 1002 decodes 3 bits from the internal packet destination address (DST). Alternatively, the global store 1004 is accessed by the switch 1002 decoding a global store access command. Any conflicts for output from the 9 from 6 switch 1002 are resolved with a conventional priority and round robin scheme. The connection from the switching network 608(1) always has highest priority. Of the 9 outputs 1010(1-9) of the 9 from 6 switch 1002, eight are connected to octants of processing element queues. An exemplary octant is designated by reference numeral 1008. Each of eight outputs 1010(1) -1010(8) are connected to an individual octant of this type. Each octant includes eight processing element queues. Each queue is 16 packets deep and includes busy/full logic and short circuits for empty queues. Each octant has only one input (from the 9-from-6 switch) and one output, and enables one read and one write to occur simultaneously.

Each cluster controller 602(1) -602(32) further includes 32 Processing Element Ports (PEPs) 1012(1)-1012(32). Each processing element port includes subports to interface with two processing

elements. Each subport includes a two byte output port connected to a corresponding one of the processing element input busses 612(1-64) and a one byte input port connected to the corresponding one of the processing element output busses 614(1-64) for each of two processing elements. The output of each queue is bused to all four PEPs (for eight processing elements) in the octant. The PEPs use address decoding to ingate only those packets which are addressed to the appropriate processing element. Each PEP includes a packet buffer for the output port with logic to signal to the octant queues when the buffer is empty.

Each of the eight octants operates independently, serving one of its eight PEP buffers one quintword each cycle, if a packet is available. From the PEPs, the packet is sent across the appropriate processing element input bus to the addressed processing element, two bytes at a time. The asymmetry of the input and output buses (one versus two bytes) helps to prevent queue full conditions.

In the inward direction (i.e. from the processing elements), one byte of data comes across one the input buses from a processing element into the corresponding processing element port (i.e. the PEP to which the PE is connected). From the processing element port, the incoming byte of data is routed directly into a port of an assembly buffer 1006 which takes in successive bytes and forms a quintword packet. The assembly buffer has 64 slots (quintword memory locations) 1014(1)-1014(64). In other words, there is one slot in the assembly buffer for each processing element, each operating independently and having its own byte counting and busy logic (not shown).

The assembly buffer slots are arranged into four columns. Each column has its own round robin logic to select one slot of those which are complete. Each cycle of the network clock, one quintword packet from one slot in each column can be outgated. The outgated packets go to the 9-from-6 switch 1002 and the 1-of-5 selector 1016. A fifth input to the 1 of 5 selector 1016 comes from the global store 1004. The 1-of-5 selector will, based on address and round robin logic, takes one packet which needs to be routed through the switch network 606 and send it on its way. Packets which are not successfully gated through either the 1-of-5 selector or the 9-of-6 switch remain in their slots to be selected the next time the round robin algorithm allows.

An example of the operation of the cluster controller, under a uniform distribution of messages, is as follows:

One input from the connected processing elements, a byte per cycle, is read into each of the assembly buffers. Five quintword packets per cycle

can be outgated to the 1-of-5 selector, so that one quintword per cycle is sent to another cluster controller.

On the output to PE direction, up to 6 quintword packets can be gated to up to 9 destinations, with queueing. Assuming a 5ns cycle of the cluster controller, with a 10ns cycle on the input and output to PE buses, the cluster controller can input 6.4 GB/sec from the PEs (100 MB/sec/PE). The assembly buffers and global memory can output 12.8 GB/sec, up to 3.2 GB/sec of which can go to other cluster controllers. Up to 19.2 GB/sec may enter into the output queues, and the output queues themselves can dispatch up to 28.8 GB/sec to the PEPs and Global Store. The PEPs each can deliver 200 MB/sec to their respective PEs, which aggregated would allow up to 12.8 GB/sec to flow out of the cluster controller to the PEs. While these are peak numbers, they show that the design is biased to allow a steady stream of 3.2 GB/sec to flow from PEs to other clusters, and up to 12.8 GB/sec back out to the PEs. Again, the design is biased to prevent queues from filling and creating contention upstream in the switch.

Fig. 12 shows a preferred embodiment of the processing elements 604(1)-604(2048) of Fig. 6. It should be understood that the present multiprocessor could use other types of processors as processing elements. The central processor 1202 of the processing element is preferably a state of the art RISC microprocessor. It is connected, in a conventional manner, to the processor cache 1204 which gives fast access time to instructions and data. The bus from the cache 1204 ties into a DMA controller 1206. The DMA controller 1206 provides the cache 1204 bidirectional ports to each of the switch buffer 1208 and the main processing element storage 1210. The switch buffer 1208 is an input/output buffer which handles the data and protocols to and from the cluster controller. The cluster controller connects to the processing element through the switch buffer 1208 by way of two unidirectional ports connected to individual busses 1212, 1214. The first unidirectional port handles incoming traffic from the cluster controller to the processing element while the second unidirectional port handles outgoing traffic from the processing element to the cluster controller.

Fig. 13 is a more detailed diagram of the DMA controller 1206 of Fig. 12. To process incoming messages, a Quintword Assembly Buffer 1302 takes 2 bytes of data at a time from the cluster controller to processing element bus 1212 and reassembles the packet. The ECC logic 1304 checks and restores the integrity of the data as well checks whether the packet arrived at the proper destination.

Once the data integrity is verified or corrected

and it is determined that the packet has arrived at its proper destination, the Input Message Control Logic 1308 places the data on a queue in the PE storage 1210. This task is accomplished by a Storage Arbitration Controller 1310, which can handle multiple requests for the PE storage 1210 and can resolve any storage conflicts. The Input Message Control Logic 1308 then signals the PE microprocessor 1202 that a message is available.

When the PE microprocessor 1202 wishes to send a message to another PE, it first enqueues the message on a destination queue in the PE storage 1210. The microprocessor 1202 then signals the Output Message Control 1312 that a message is ready. It does this by doing a "store" operation to a fixed address. This address does not exist in the PE storage 1210 but is decoded by the Storage Arbitration Control 1310 as a special signal. The data for the "store" operation points to the destination queue in PE storage 1210.

Before being sent to the cluster controller, each message in the destination queue is provided with a header. The headers are kept locally in the DMA controller 1206 in the destination PE Q-header array 1314. The message header specifies the total length of the message in bytes (up to 4096), the id of the PE to which the message is to be sent (15-bit DST id), and the id of this sending PE (15-bit SRC id).

To achieve high switch bandwidth, the DMA controller interleaves packets from multiple messages, rather than send the messages sequentially. However, all messages from one processing element to another specific processing element are sent in order. The switch design ensures that the packets received by a processing element from another specific processing element are received in the same order in which they were sent. The Output Message Control Logic pre-fetches all or portions of the top message for the various destination into the Output Message Buffer 1316. From the Output Message Buffer 1316, the data is taken, one quintword at a time into the Quintword Disassembly Buffer 1318 where it is sent, a byte at a time, across to the cluster controller.

As a further function, the DMA controller 1206 also generates a nine bit SEC/DED Error Correcting Code (ECC) for each packet prior to transmission.

The error correction function of the present system will now be described in more detail. As previously explained, as message packets arrive at a processing element, the DMA controller 1206 applies the ECC, and then performs the function specified by the packet command field. If the ECC indicates that a single bit error occurred in the DST id of the received packet, then the packet should have gone to some other processing element, so

the DMA controller 1206 corrects the DST id and retransmits the packet to the correct processing element. Where the cluster network is configured with a host processor, the DMA controller 1206 also reports this error event to a host processor service subsystem. This is accomplished by generating an interruption to software on the host processor, which reports the error to the service subsystem under the control of a thresholding algorithm.

While ECC is generated in the sending processing element and applied in the receiving processing element, parity checking is also performed every time a packet enters or leaves a TCM, and upon receipt by the destination processing element. Thus, correctable errors are detected and can be reported to the service system as soon as they begin to occur.

The self correcting error handling of the present system will be better understood by reference to Fig. 6. We will assume, for example, that there is a cabling problem between cluster 602(1) and the 32X32 switch network 606, that will cause a hard error in the destination address field of an incoming packet. We will further assume that the incoming packet was intended for processing element 604(3) on cluster controller 602(1) but instead, due to the hard error, arrives as processing element 604(1) on the same cluster controller.

The receiving processing element 604(1) will receive the incoming packet by way of the 9-6 switch and a PEP output bus. Once the packet is received, the processing element will correct the destination field error (using the ECC), and resend the packet on the cluster controller 602(1) to the correct PE 604(1) by way of the PEP input bus and the assembly buffer. Since the packet will no longer travel the path of the problem connection, the hard error will not be repeated for this packet.

A similar procedure ensures correction of many errors where an incorrect destination address is caused on the bus from the cluster controllers to the switch network 606. It will be noted that each cluster has a separate input bus and output bus. Therefore, if the destination address of an outgoing packet is altered due to a misconnection on the output side of the bus and a packet is sent to the wrong cluster controller, the path between the correct cluster controller and the receiving/correcting cluster controller will completely differ from the path between the originating processor and the receiving/correcting processor.

The switch network 606 itself also includes error correction logic. Therefore, if a packet is routed to a non-present or non-operational processing element, the switch will reverse the source and destination fields and send the packet back to the sender with an error indication.

Fig. 11 shows a preferred embodiment for a packet format used with the system of Fig. 6. Each packet is 180 bits wide and includes a 5 bit command field (CMD), an 8 bit sequence number field (SEQ), a 15 bit destination address field (DST), a 15 bit source address field (SRC), a 128 bit data field and a 9 bit error correction code (ECC).

The command field (CMD) includes a five bit command that tells the cluster controller and the receiving processing element how to handle the packet. The sequence number field (SEQ) includes an 8 bit packet sequence number sequentially assigned by the originating (source) processing element. The sequence number enables the receiving system to identify which packet number of the total packet count in the message has been received.

The destination address field (DST) includes a fifteen bit destination processing element number. The destination field is used by the switch and cluster controller to self route the packet and by the receiving (destination) processing element to verify that the packet has been routed to the proper address.

The source address field (SRC) includes a fifteen bit originating (source) processing element number. The source field is used by the switch and cluster controller to return the packet to the source in a case where an inoperable or non-present processing element number appears in the destination address field (DST) field, and by the receiving (destination) processing element to properly address any response to the message or command.

The data field (DATA) includes 128 bits of information. The type of information in the data field is defined by the command field (CMD).

The ECC Field (ECC) includes an SEC/DED (Single Error Correct/Double Error Detect) error correction code.

For message header packets, the sequence field specifies the total length of the message, and the DMA controller allocates a message buffer of this length in the PE local memory, writes the initial quadword of data into the message buffer, and sets local hardware pointer, length and sequence registers if there will be more packets of data for this message. It also constructs the message header in memory, which includes the message length, DST id and SRC id.

For message body packets, the sequence number field is checked against the sequence register to verify that packets are arriving in order, and each quadword of data is added to the message buffer. When the message has been completely received it is enqueued on a queue in local memory, known as the IN\_QUEUE, for processing by the local processor. If the IN\_QUEUE had been empty prior to the addition of this message, then an interruption is generated to the local processor

to notify it of pending work.

For storage access command packets, the DMA controller performs the required fetch or store operation to the PE local memory (transferring a doubleword of data), and for fetches a response packet is constructed by reversing the SRC and DST id fields, and then sent on through the switch to return the requested doubleword of data.

Packets that contain global storage access commands are handled in the cluster controller in the same way that local storage access commands are handled by the DMA controllers. In both cases, the memory operations are autonomous, and include a compare-and-swap capability.

Fig. 14 depicts a preferred layout of a processing element/cluster board. In terms of physical layout, a cluster preferably comprises a multilayer circuit board 1400 on which up to 64 processing element cards (i.e. circuit boards which each embody a processing element) are mounted directly, and at least one cluster controller thermal conduction module (TCM) 1402. Each cluster controller handles local message passing within the cluster, and connects to the switch network 606.

Fig. 15 shows a preferred system frame layout with four clusters in each of eight frames 1502-1516. The switch network thermal conduction modules are preferably embodied in central frames 1518-1524. The Host Adapter 1700 (Fig. 17) can reside in any one of the switch network frames 1502-1516. For availability and configurability reasons, an additional Host Adapter 1700 can be provided in another one of the switch network frames 1502-1516.

Fig. 16 shows a preferred layout for a processing element card 1600, including the high performance RISC microprocessor 1202, the optional database accelerator 1602, the DMA controller 1206, and the local memory 1210. The processing element cards 1600 have twice as many pins as can be connected to the cluster controller TCM. Therefore, a second set of PE buses (a second "PE port") is brought off the processing card and onto the mother board (the TCM mother board) where it is routed to the second (spare) cluster controller TCM position (1404, Fig. 14). This allows for future expansion: as CMOS densities continue to improve, a second PE could be packaged per card, and duplicate cluster controller and switch network TCM's could be plugged into the pre-wired boards, doubling the size of the system to 4096 PEs. Alternatively, with the optional cluster controller and switch network TCM's plugged, each PE could use two PE ports to the cluster controller, either for higher bandwidth or for improved fault tolerance.

The above-described system can be built as a stand-alone multiprocessing system, as a stand-

alone database processor or employed as a coprocessor to a traditional mainframe host. In the latter case, the host system would provide the front-end MVS/DB2 system functions, including session management, transaction processing, database locking and recovery. The present multiprocessor system could also be employed as a back-end system to offload and accelerate the read-only complex query processing functions from the host.

Many modifications and variations which can be made without departing from the scope and spirit of the invention will now occur to those of skill in the art. It should be thus understood, that the present description of the system provided as an example and not as a limitation.

### Claims

1. A multiprocessing system having at least three nodes, comprising:
  - a first node (604) comprising means to transmit data comprising destination identifying information;
  - a second node (604) coupled to said first node (604), said second node comprising means for receiving said data along a first path, means (1206) for detecting and correcting an error in said destination identifying information so as to form corrected destination identifying information and means for rerouting said data, along a second independent path, to a third node (604) identified by said corrected destination identifying information.
2. The multiprocessing system of claim 1 wherein said first, second and third nodes are processors and wherein said processors are each coupled to a self routing switch (606) by way of independent input (608, 612) and output (610, 614) data paths.
3. The multiprocessing system of claim 2 wherein said data is packetized and comprises a destination field including said destination identifying information and a source field identifying a source processor.
4. The multiprocessing system of claim 2 wherein said switch (606) comprises means for detecting when said destination field identifies a non-present processor, for reversing said source and destination fields and for rerouting said data to said source processor.
5. A method of error recovery in a multiprocessor system connected by a switching network, wherein a first processor (604) in said system transmits a data packet having an address field comprising an address of a second processor (604) in said system, comprising the steps of:
  - transmitting said packet from said first (604) processor to said switching network (606) by way of a first path;
  - decoding said address field in said transmitted packet at said switching network (606);
  - routing said packet by way of a first path from said switching network (606) to a third processor (604) in said system designated by said decoding;
  - detecting at said third processor (604), an error in said address field of said packet;
  - correcting said error at said third processor (604) to form a corrected address in said address field;
  - retransmitting said packet having said corrected address from said third processor (604) to said switching network (606) by way of a second path;
  - decoding said address field in said retransmitted packet at said switching network (606); and
  - routing said retransmitted packet from said switching network to said second processor (604) by way of a third path.
6. The method of claim 5 comprising the further steps of:
  - determining at said switch (606) if a decoded address corresponds to a non-operable processor; and
  - when said determining determines that said decoded address corresponds to said non-operable processor, causing said switch (606) to exchange said source and destination fields in said packet and returning said packet to said first processor (604) by way of a fourth path;
7. A cluster controller (602) for use in a multiprocessor system comprising a plurality of processor clusters coupled by way of a switching network, said cluster controller (602) comprising:
  - switching means (1002), connected to receive

packets from said switching network, for distributing said packets from said switching network in accordance with a destination address;

global storage means (1004) for storing data, said global storage means being connected to receive said packets from said switching means (1002);

queue means (1008) for buffering packet flow to a plurality of processors, said queue means comprising a plurality of packet queues associated with each of said processors;

a plurality of first busses, each of said first busses being connected to an output port of said switching means and an input port of one of said packet queues, said first busses having a first number of bits;

a plurality of processing element port means (612, 614) for transferring data between said cluster controller and said processors;

a plurality of second busses, each of said second busses being connected to an output port of one of said packet queues and an input port (612) of one of said processing element port means,

assembly buffer means (1006) for assembling data from said processors into packets, said assembly buffer means (1006) comprising one assembly buffer (1014) for each of said processors and round robin means for selecting an assembled packet to be output, said assembly buffer means (1006) being connected to receive said data from said processing element ports; and

selector means (1016) for selecting one packet to be sent to said switching network (606), said selector means being connected to receive packets from said assembly buffer means (1006) and said global store means (1004).

8. The system of claim 7 wherein said selector means (1016) further comprises selector means for outputting said packets in round robin fashion.

9. The system of claim 7 wherein said second busses have a second number of bits larger than said first number of bits;

10. A cluster connected multiprocessing system comprising:

5

10

15

20

25

30

35

40

45

50

55

11

a first plurality of processors (604), wherein each of said processing elements in said first plurality comprises a local memory;

a second plurality of processors (604), wherein each of said processing elements in said second plurality comprises a local memory;

first cluster controller means (602) connected to receive first data from said first plurality of processors (604), for assembling said first data into packets comprising a source field, a destination field and a command field, and for outputting said first plurality of packets;

second cluster controller means (602) connected to receive second data from said second plurality of processors (604), for assembling said second data into packets comprising a source field, a destination field and a command field, and for outputting said packets; and

switching network means (606) connected to receive said packets from said first and second cluster controller means (602), for decoding said destination field and for determining which one of said cluster controller means is connected to an addressed processors corresponding to said decoded destination field and for routing said packets to said one of said cluster controller means (602).

11. The system of claim 10 wherein said first and second cluster controller means (602) each comprise means for outputting said packets in round robin fashion.

12. The system of claim 10 wherein each of said processors (604) in said first and second plurality comprises means for providing direct access to said local memory by every other processor in said first and second plurality.

13. The system of claim 10, further comprising:

host adaptor means (1700), for coupling a host processor to said switching network means, said host adaptor means comprising:

means (1714) for receiving a set of commands from said host processor; and

means (1710) for distributing said commands among a plurality of said processors.

14. The system of claim 13 wherein said host adaptor means further comprises:

means (1706) for translating first memory addresses from said host processor to a band of second memory addresses in a local memory within each of said plurality of said processors.

15. The system of claim 14 wherein each of said processors comprises a general purpose processor and a database accelerator.

16. The system of claim 15 wherein at least one of said database accelerators is a sort coprocessor.

17. A cluster controller (602) for use in a multiprocessor system comprising a plurality of processing element clusters coupled by way of a switching network (606), said cluster controller (602) comprising:

switching means (1002), connected to receive packets from said switching network (606), for distributing said packets from said switching network (606) in accordance with a destination address;

queue means (1008), coupled to a plurality of said processing elements (604), for buffering packet flow to said plurality of processing elements, said queue means comprising a plurality of packet queues associated with each of said processing elements,

assembly buffer means (1006), coupled to said plurality of said processing elements (604), for assembling data from said processing elements (604) into packets, said assembly buffer means (1006) comprising one assembly buffer (1014) for each of said processing elements (604); and

selector means (1016), coupled to said assembly buffer means (1006) for selecting a packet from said assembly buffer means (1006) to be sent to said switching network (606).

18. A packet format for use in a cluster connected multiprocessing system comprising:

a command field comprising:

a first defined pattern of bits which when decoded by a cluster controller within said multiprocessing system will cause a write to a global memory within said cluster controller;

a second defined pattern of bits which identifies a packet including said command field as

carrying a message body;

a third defined pattern of bits which identifies a packet including said command field as carrying a message header;

a sequence number field for carrying any of a sequence number of a packet where said command field defines said packet as a message body, and a count of message packets to follow where said command field defines said packet as a message header;

a destination field for carrying an first address of a destination processing element in said cluster connected multiprocessing system;

a source field for carrying a second address of a source processing element in said cluster connected system;

a data field; and

an error correction code field for carrying an error correct, error detect correction code.

19. The packet format of claim 18, wherein said command field further comprises a fourth defined pattern of bits which when decoded by said cluster controller will cause a local memory in a processor connected to said cluster controller to be accessed.

FIG. 1 (PRIOR ART)

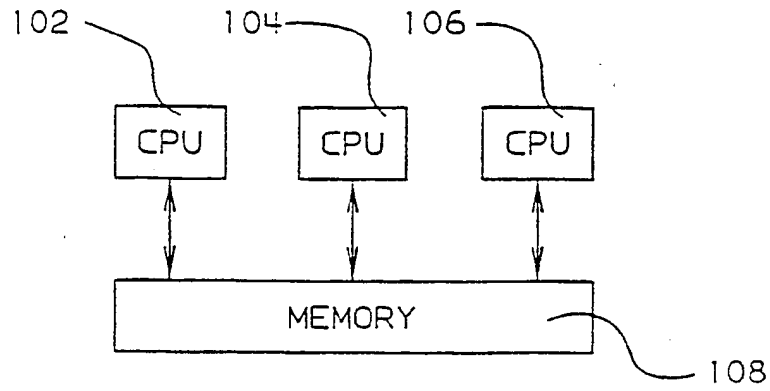


FIG. 3 (PRIOR ART)

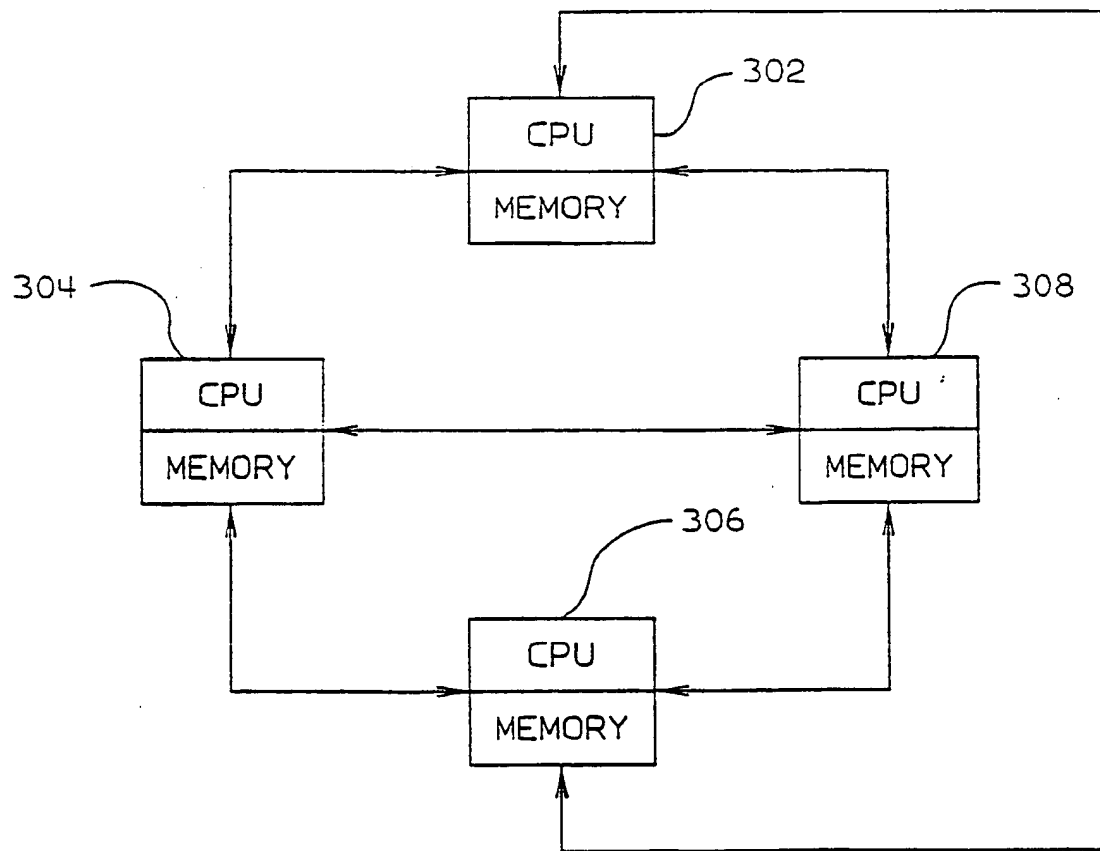


FIG. 2 (PRIOR ART)

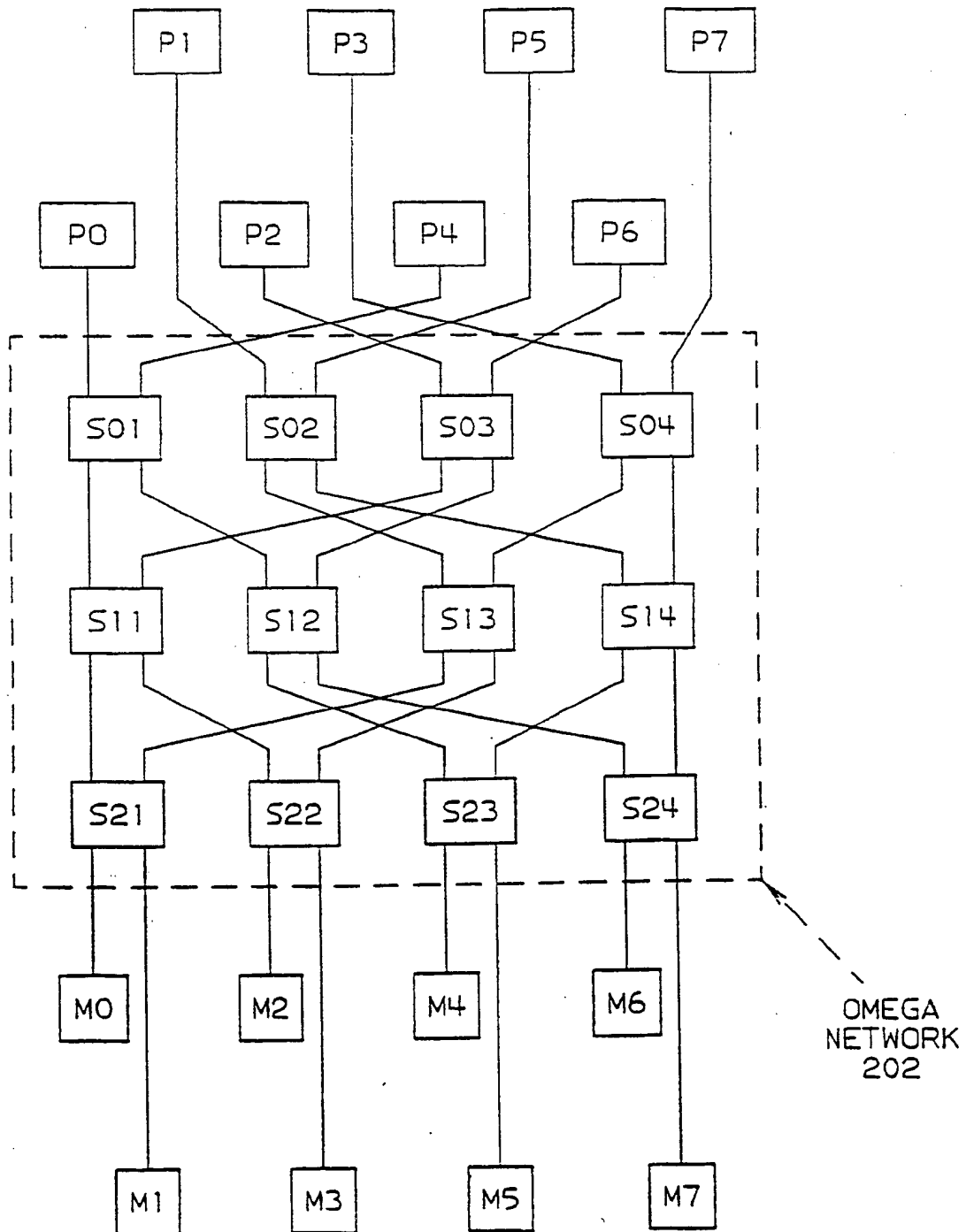


FIG. 4 (PRIOR ART)

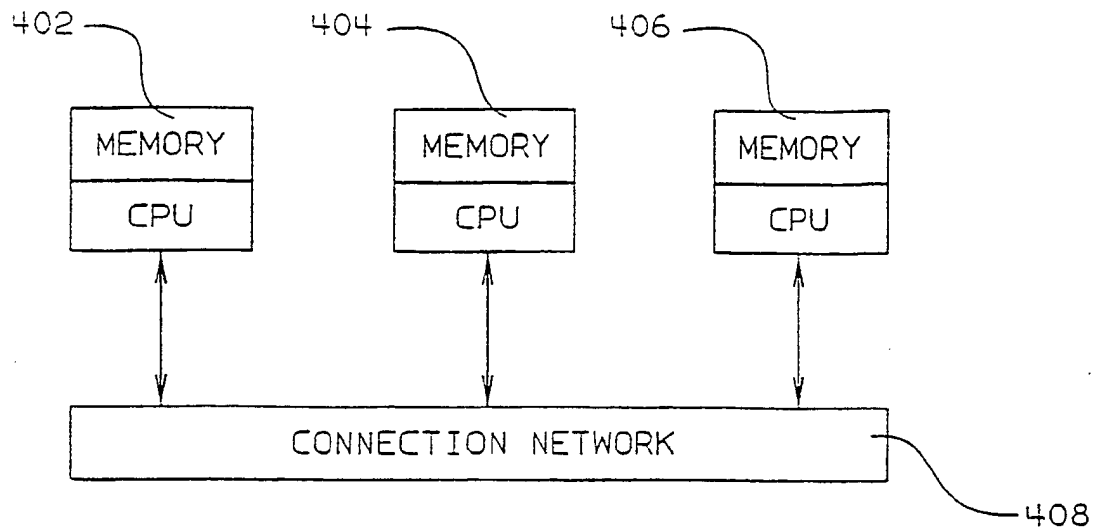


FIG. 5 (PRIOR ART)

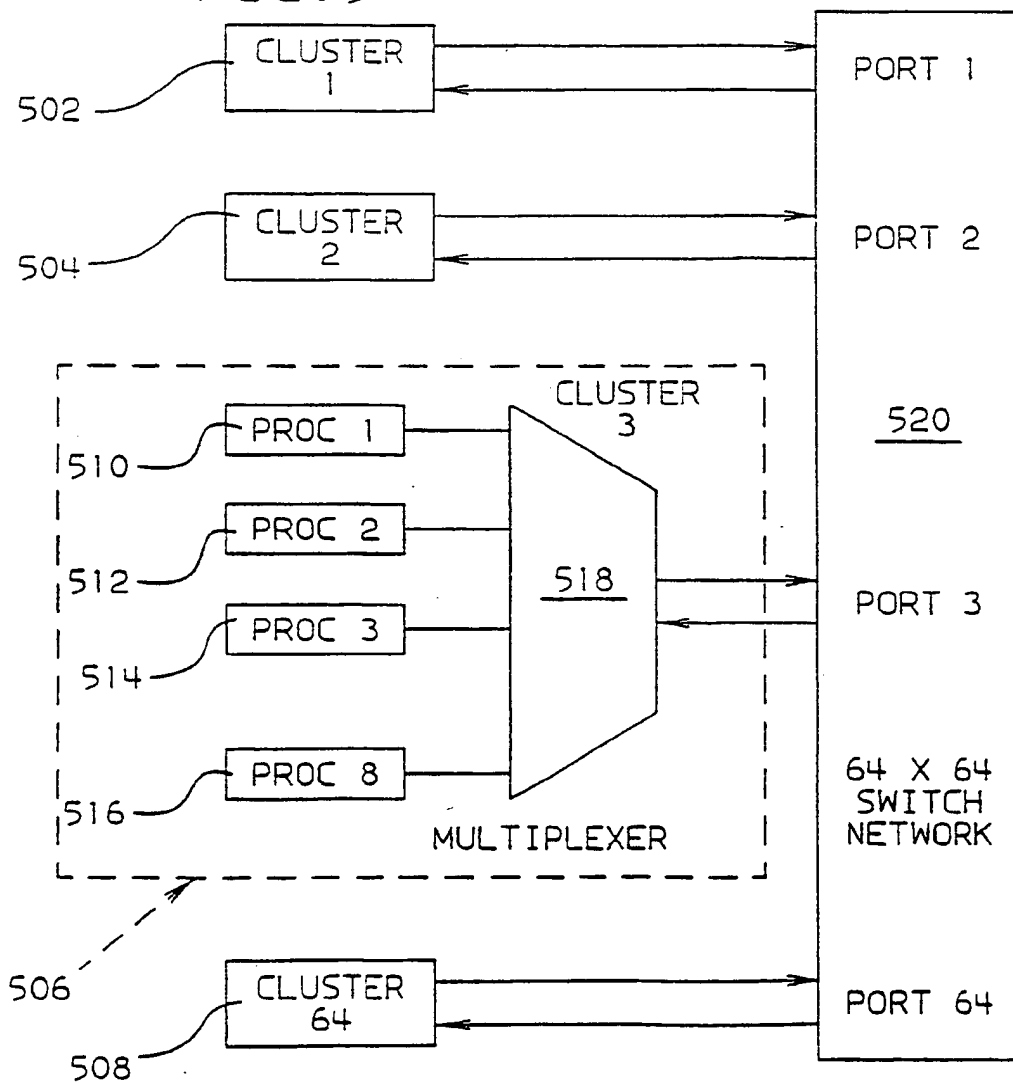


FIG. 6

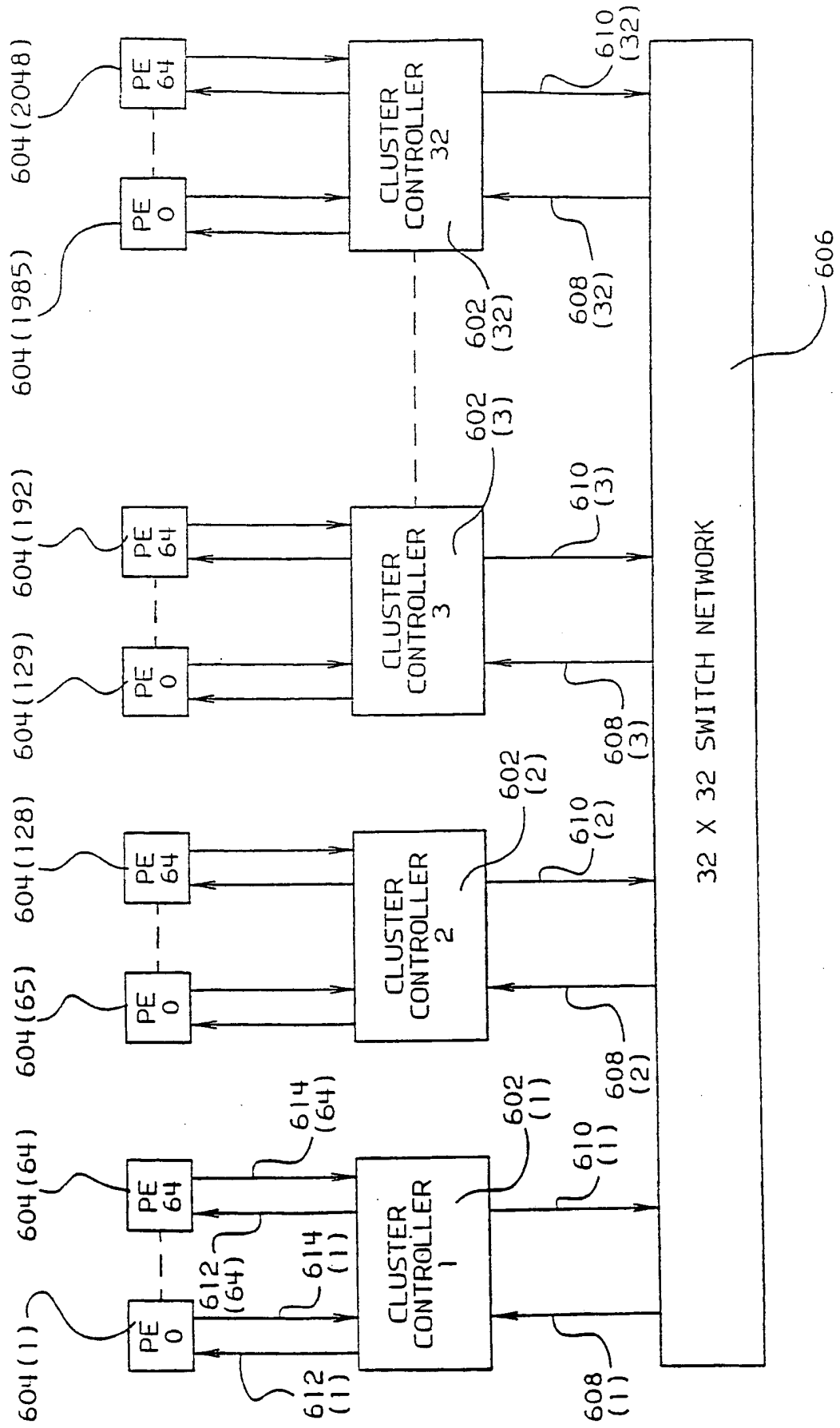


FIG. 7

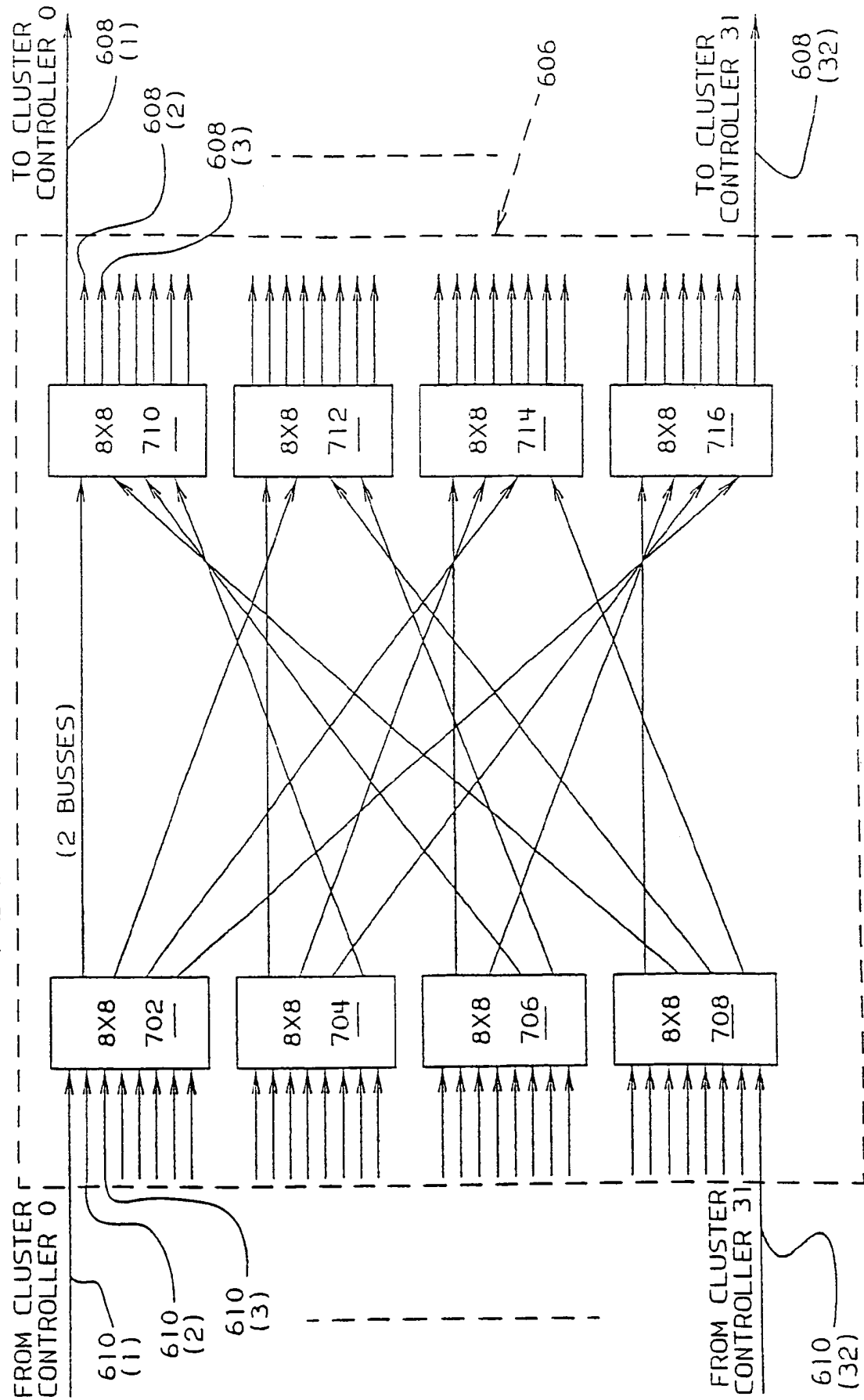


FIG.8

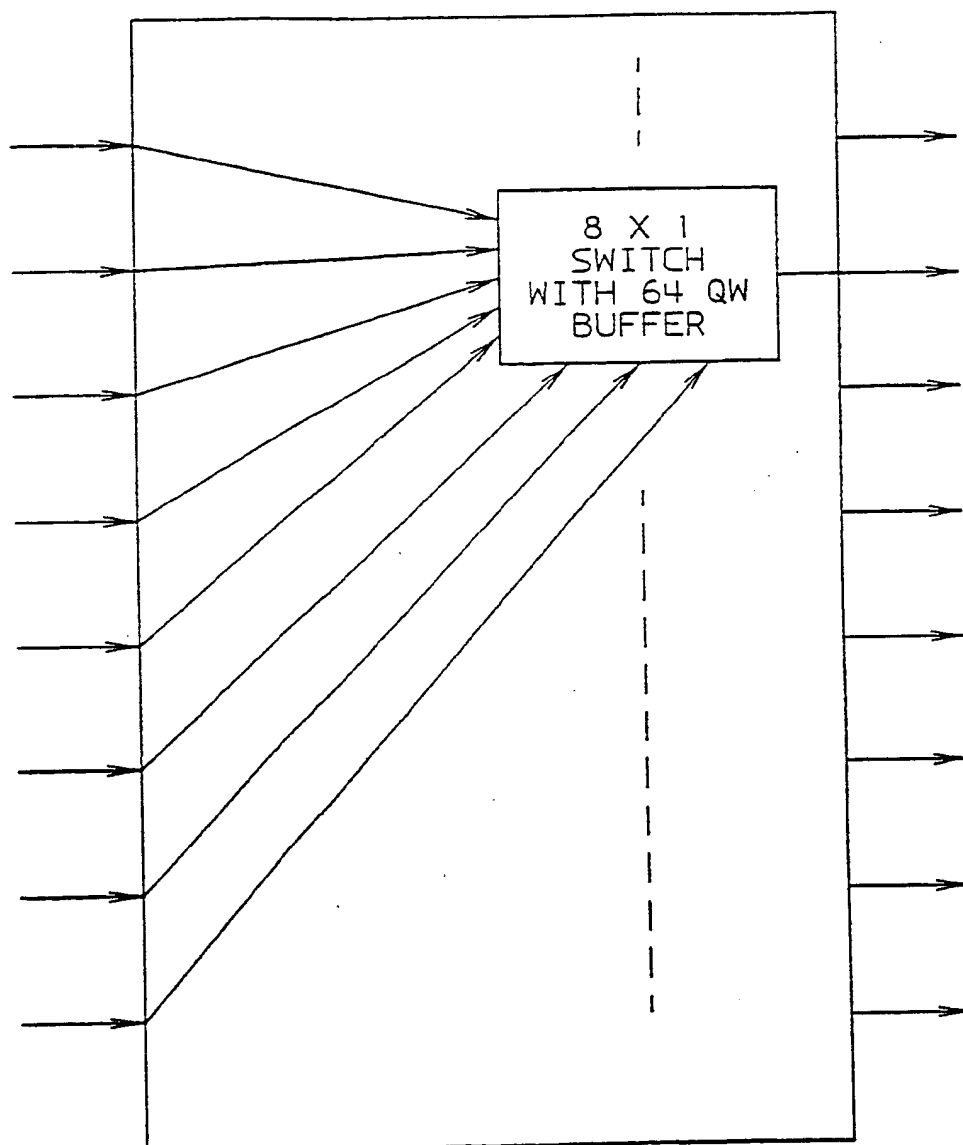
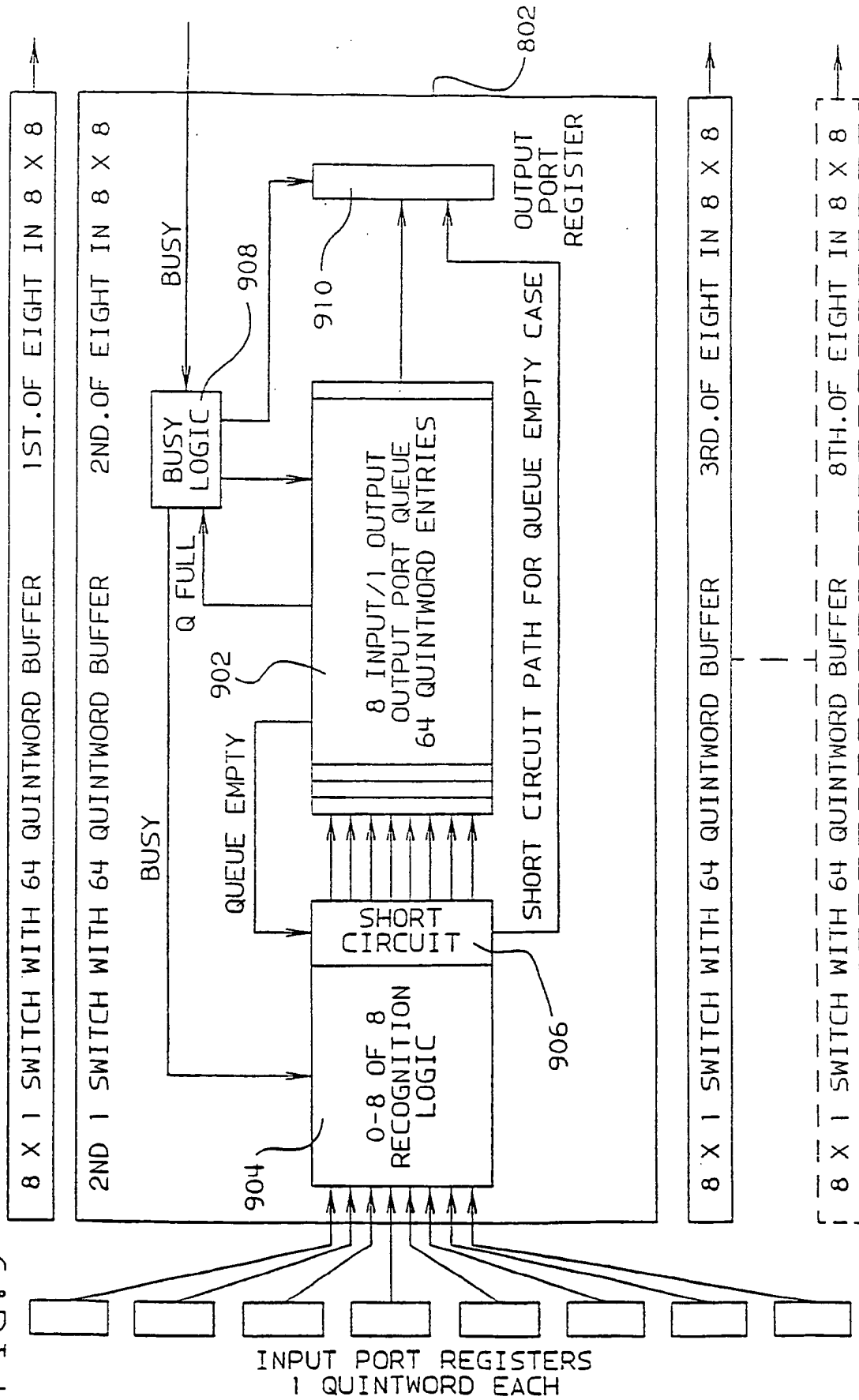


FIG. 9



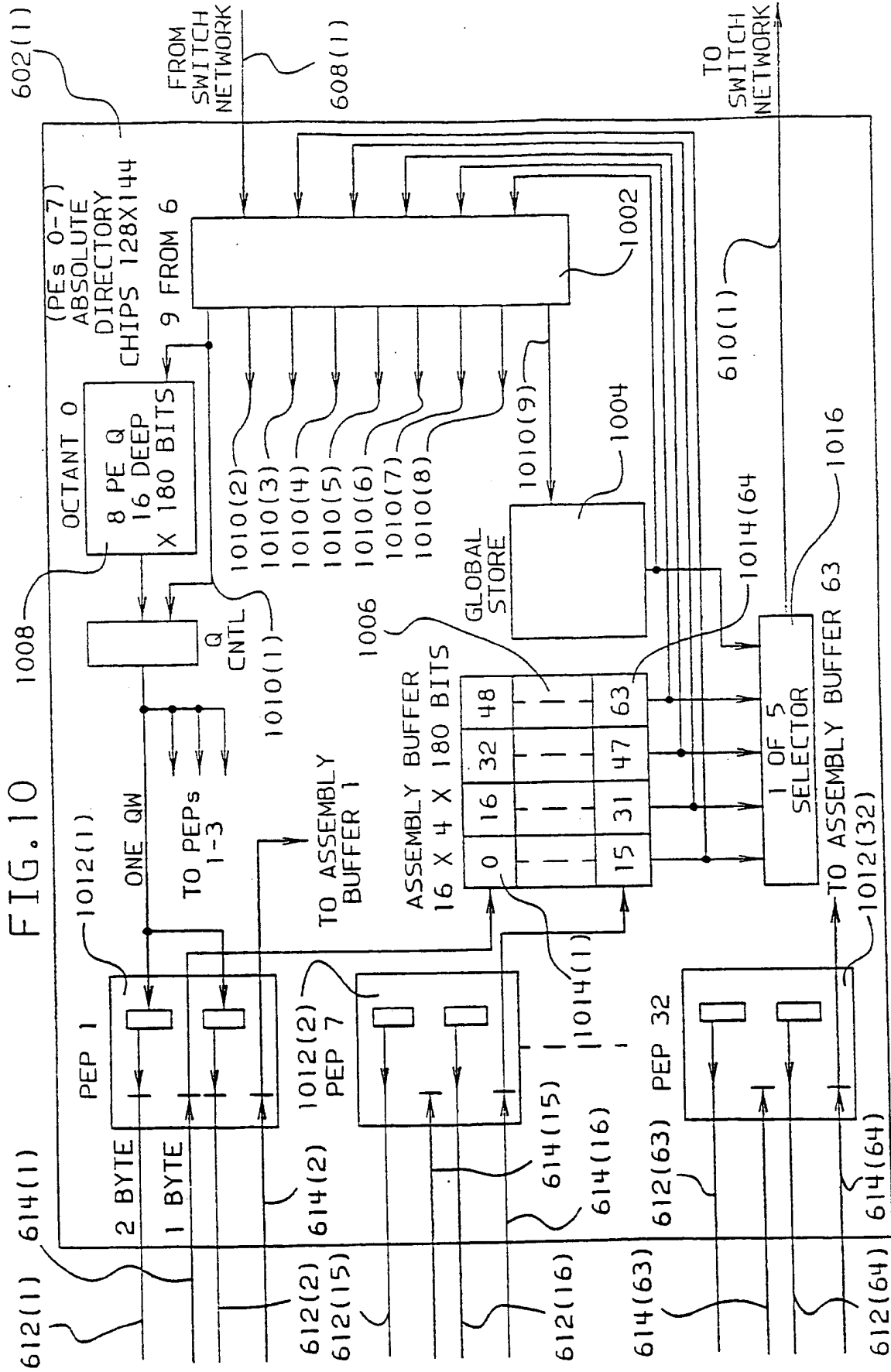


FIG. 11

## PACKET FORMAT

CMD	SEQ	DST	SRC	DATA	ECC
5	8	15	15	128	9

FIELD	SIZE	MEANING
CMD	5	COMMAND FIELD:  00001-MESSAGE HEADER;DATA FIELD HOLDS FIRST 16 BYTES OF MESSAGE;SEQ FIELD HOLDS COUNT OF MESSAGE BODY PACKETS THAT FOLLOW(TOTAL MESSAGE IS 1-256 PACKETS, 16-4096 BYTES) 00010-MESSAGE BODY 00011-CONTROL FUNCTION,DATA FIELD PROVIDES SUBCOMMAND 00100-GLOBAL STORAGE ACCESS,DATA FIELD PROVIDES OP/ADDR/DATA 00101-PE STORAGE ACCESS,DATA FIELD PROVIDES OP/ADDR/DATA 10000-GLOBAL BROADCAST TO ALL PE's 10001-BROADCAST ALL PE's ON S1 ADDRESSED BY DST,PER MASK IN DATA FIELD(0-63);DATA FIELD (64-127)HOLDS MESSAGE
SEQ	8	SEQUENCE NUMBER IN MESSAGE BODY PACKETS;IN MESSAGE HEADER PACKETS, COUNT OF MESSAGE BODY PACKETS TO FOLLOW
DST	15	DESTINATION PE NUMBER
SRC	15	SOURCE PE NUMBER
DATA	128	DATA CONTENT FOR DATA PACKETS; SUBCOMMAND FOR, CONTROL PACKETS; OPERATION TYPE,ADDRESS AND DATA FOR STORAGE ACCESSSES
ECC	9	SEC/DED ERROR CORRECTION CODE

180 BITS PER PACKET

FIG. 12

## PROCESSOR ELEMENT

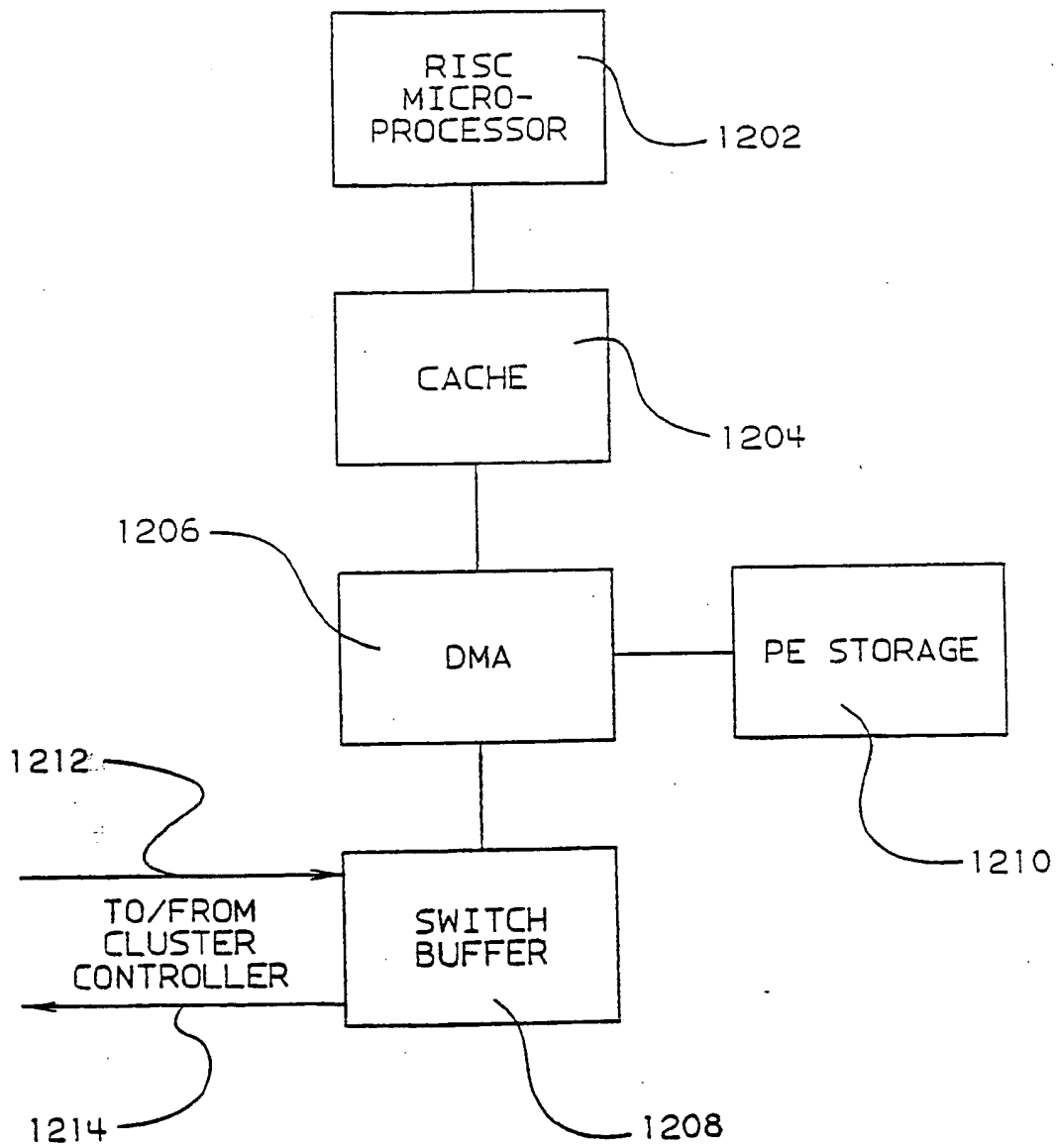


FIG. 13

DMA  
CONTROLLER

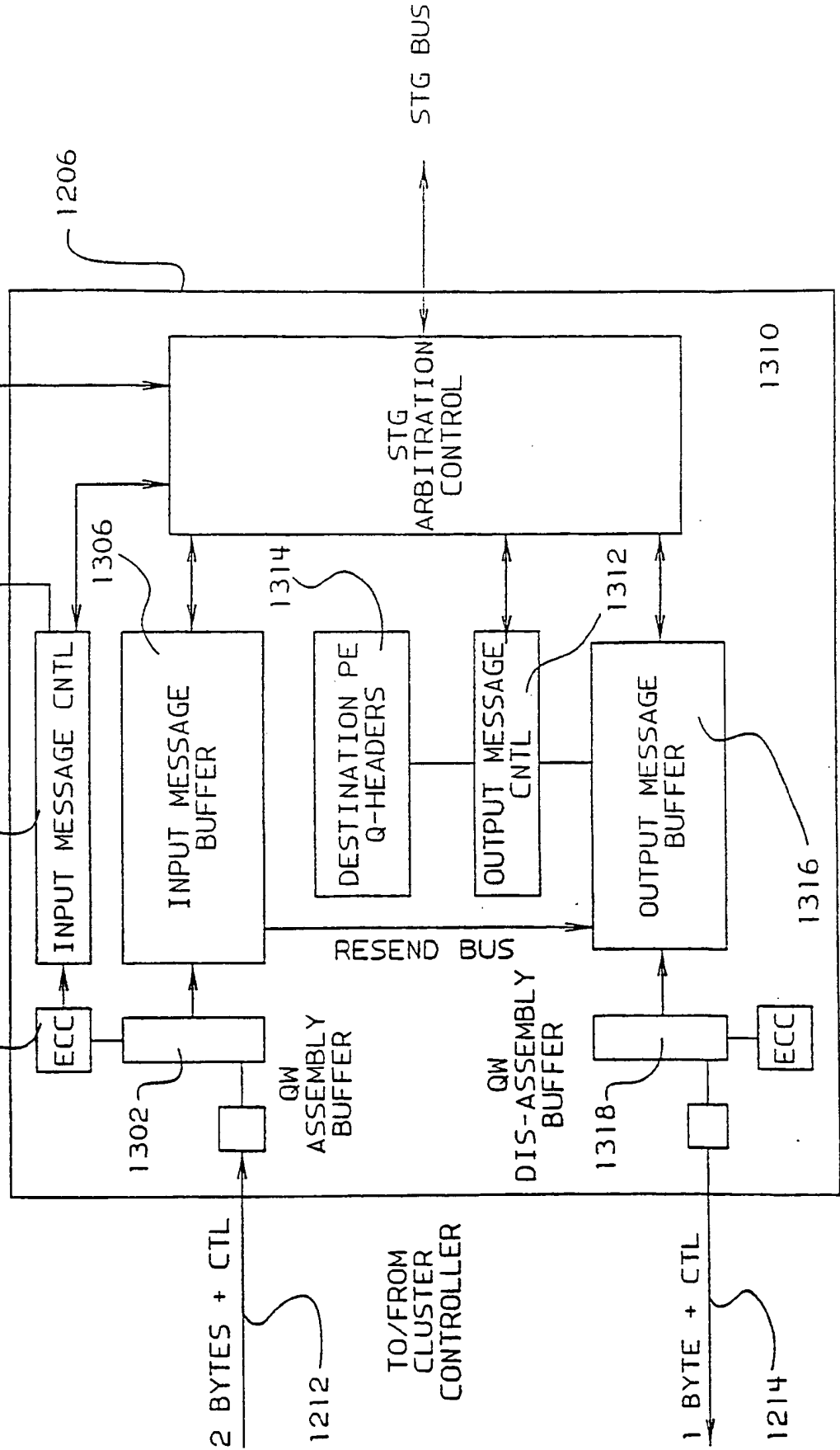


FIG. 14

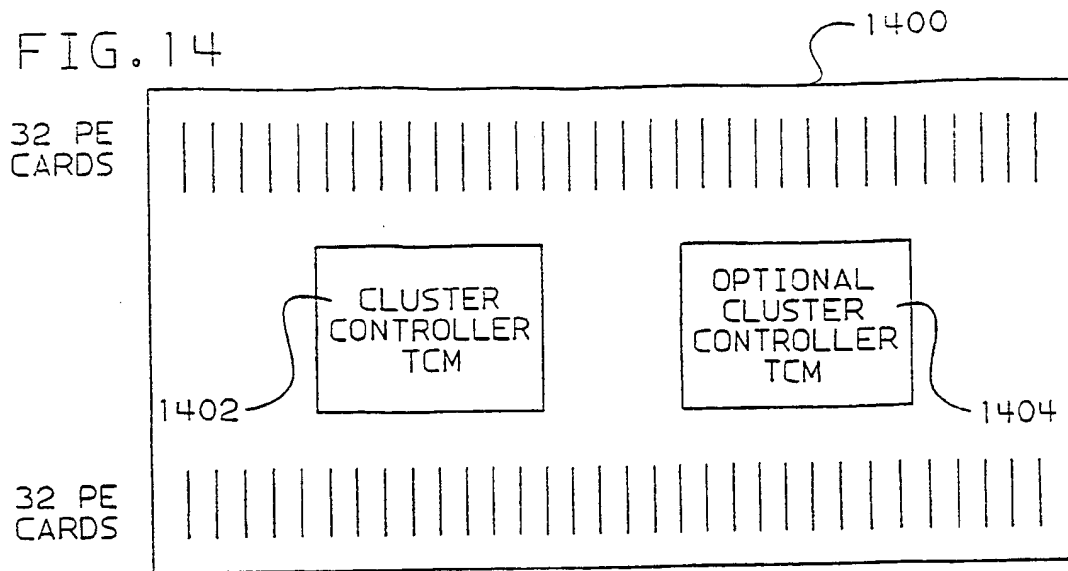
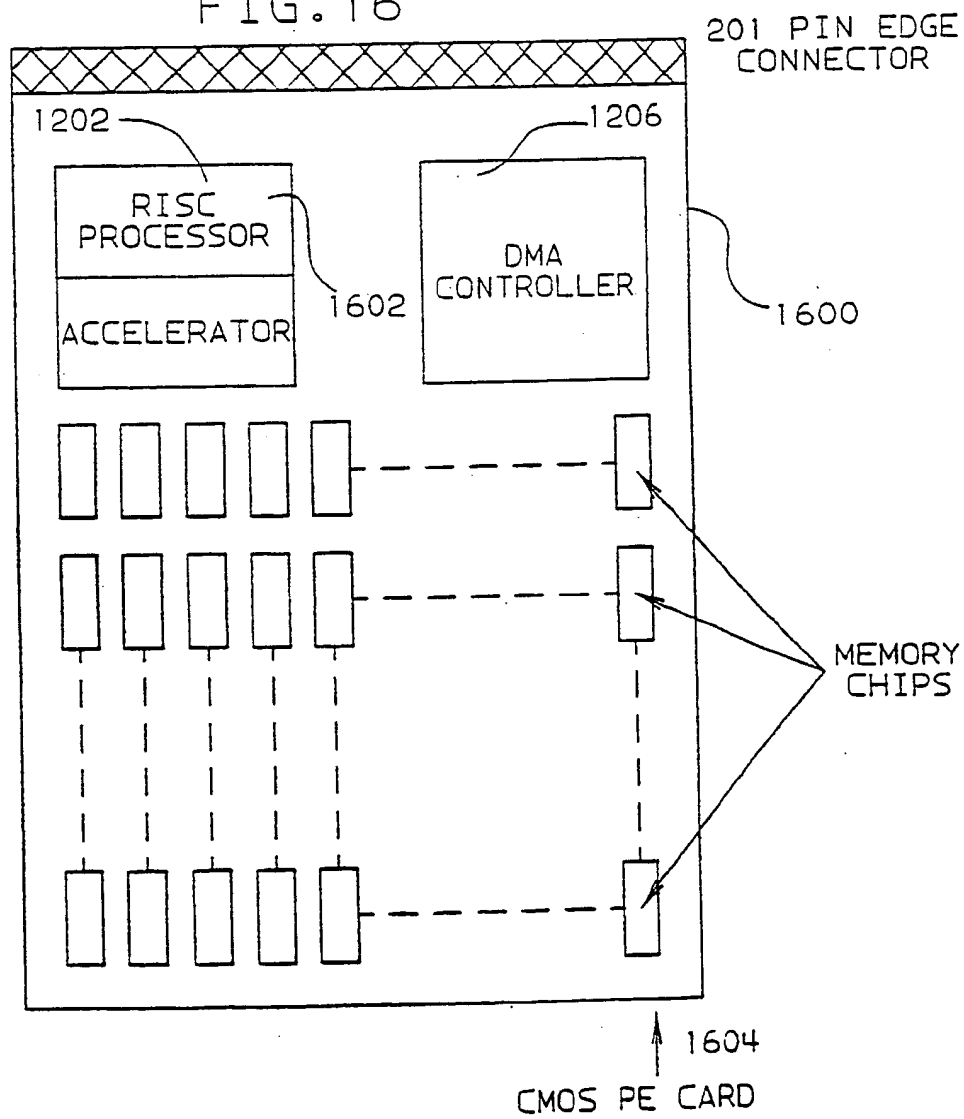
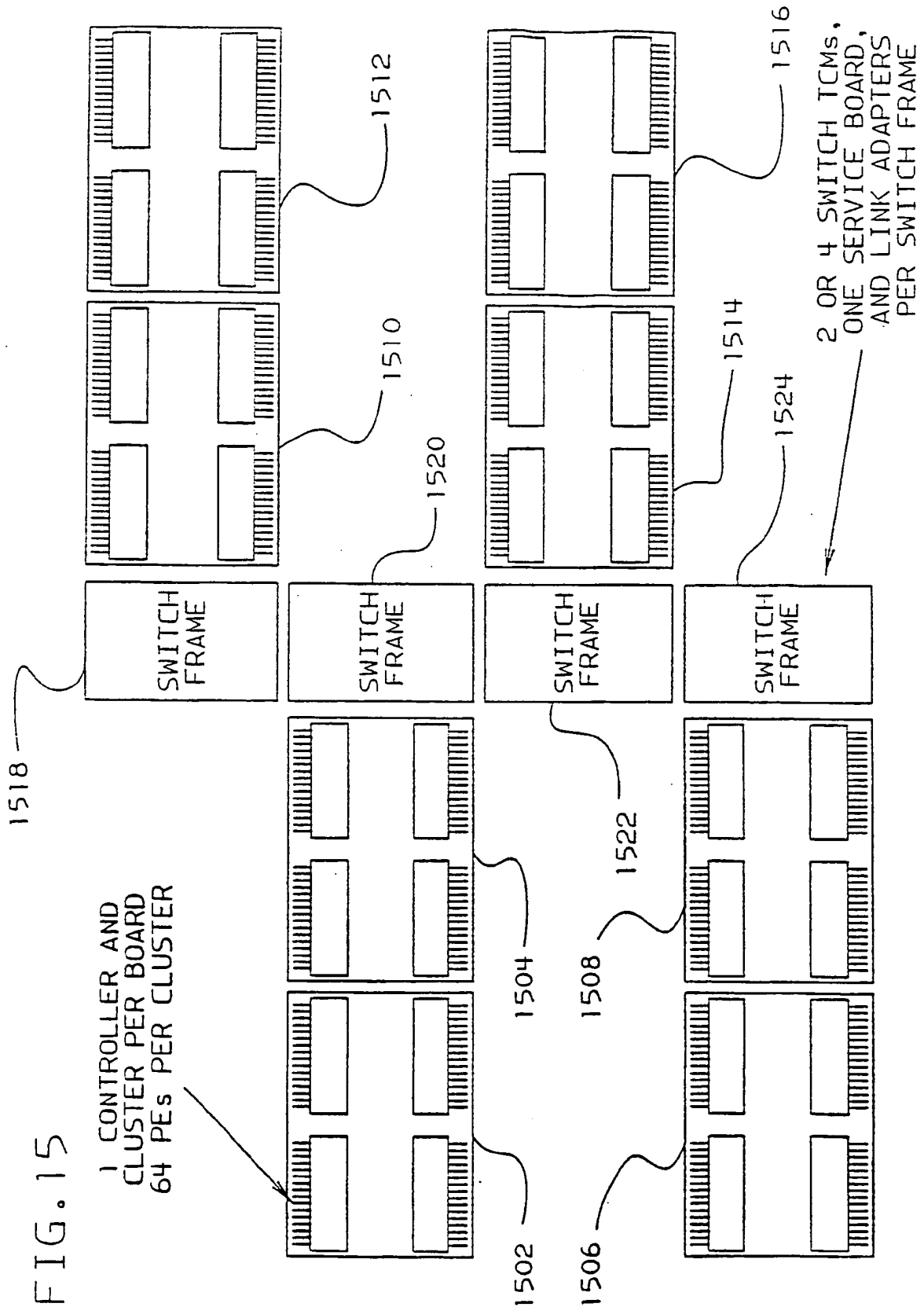
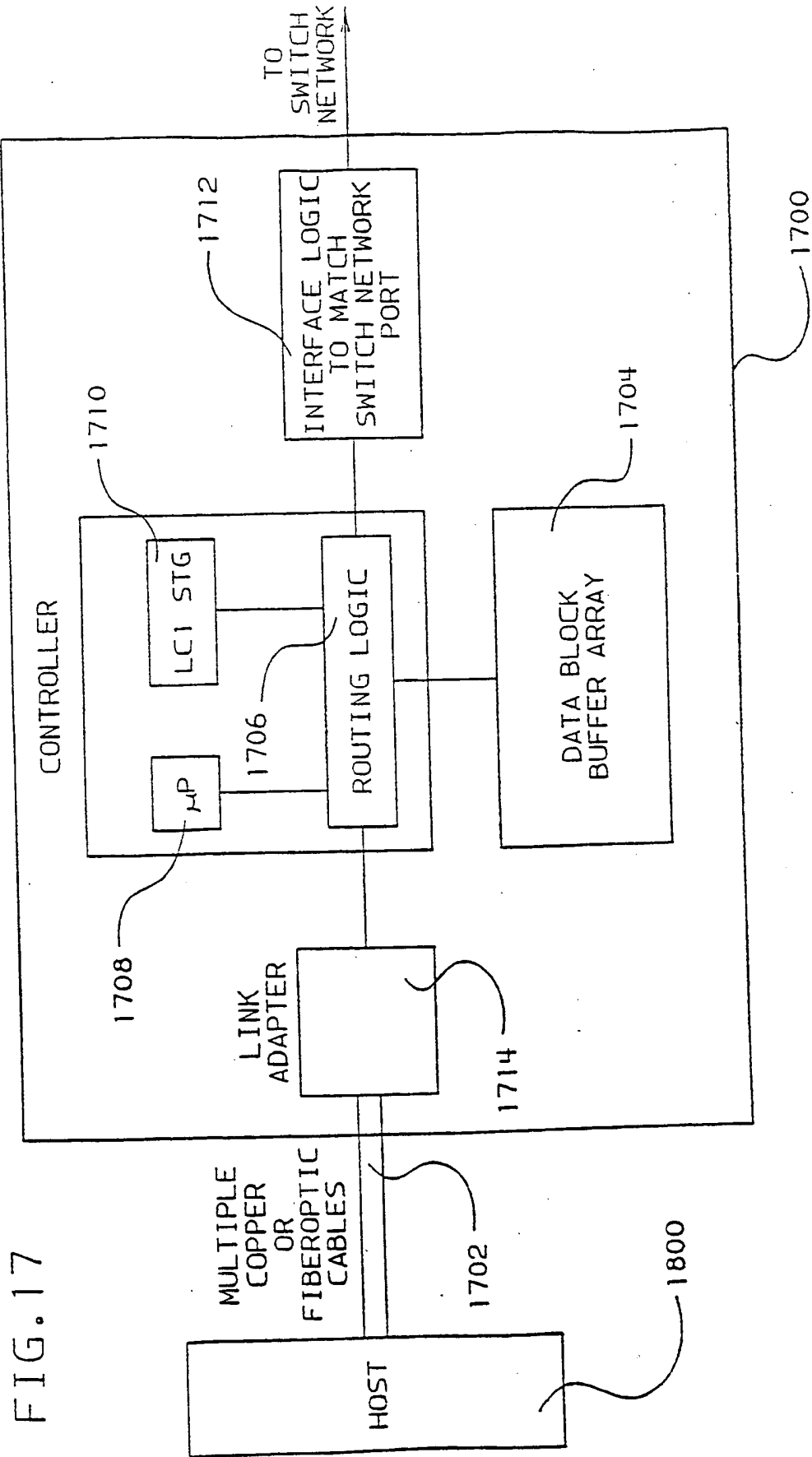


FIG. 16









Europäisches Patentamt  
European Patent Office  
Office européen des brevets



Publication number: **0 439 693 A3**

12

## EUROPEAN PATENT APPLICATION

21 Application number: **90120874.4**

51 Int. Cl.<sup>5</sup>: **G06F 11/00, G06F 11/10**

22 Date of filing: **31.10.90**

30 Priority: **02.02.90 US 474440**

43 Date of publication of application:  
**07.08.91 Bulletin 91/32**

64 Designated Contracting States:  
**DE FR GB**

68 Date of deferred publication of the search report:  
**24.06.92 Bulletin 92/26**

71 Applicant: **International Business Machines Corporation**  
**Old Orchard Road**  
**Armonk, N.Y. 10504(US)**

72 Inventor: **Baum, Richard Irwin**  
**5 Arbor Hill Drive**  
**Poughkeepsie, New York 12603(US)**  
Inventor: **Brotman, Charles H.**  
**13 Saint Annes Road**  
**Poughkeepsie, New York 12601(US)**  
Inventor: **Rymarczyk, James Walter**  
**6 Dara Lane**  
**Poughkeepsie, New York 12601(US)**

74 Representative: **Jost, Ottokarl, Dipl.-Ing.**  
**IBM Deutschland GmbH Patentwesen und**  
**Urheberrecht Schöналcher Strasse 220**  
**W-7030 Böblingen(DE)**

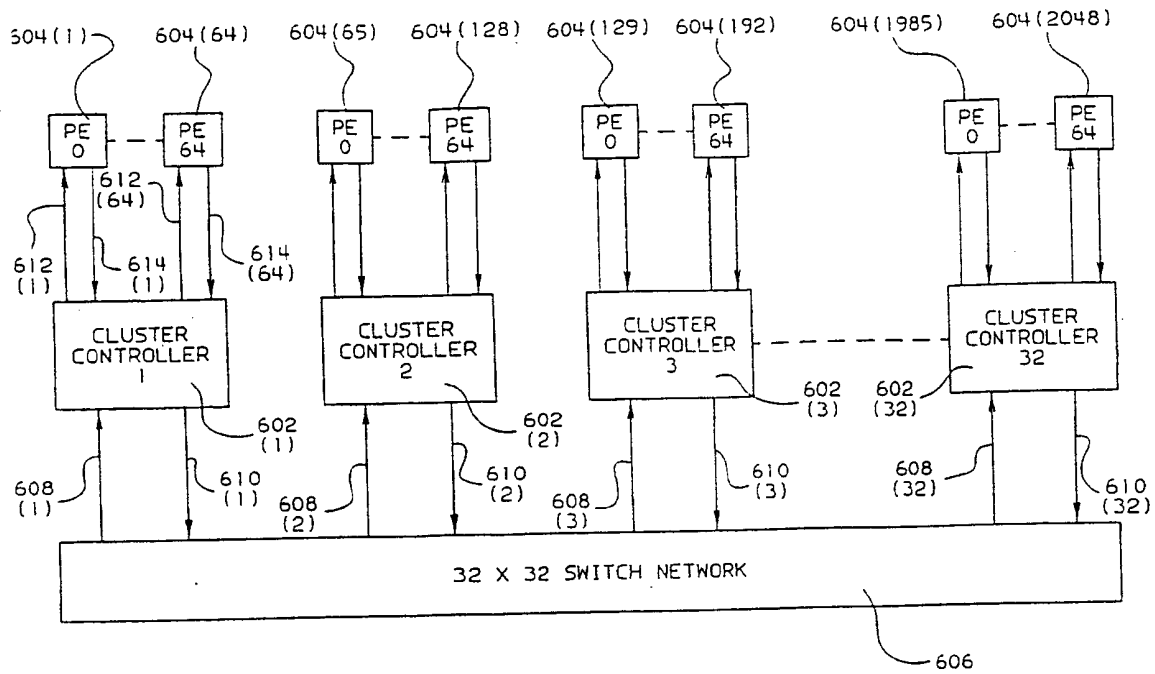
54 **Multiprocessing packet switching connection system having provision for error correction and recovery.**

57 A large number of processing elements (604) (e.g. 4096) are interconnected by means of a high bandwidth switch (606). Each processing element (604) includes one or more general purpose microprocessors (1202), a local memory (1210) and a DMA controller (1206) that sends and receives messages through the switch (606) without requiring processor intervention. The switch (606) that connects the processing elements is hierarchical and comprises a network of clusters. Sixtyfour processing elements (604) can be combined to form a cluster and and sixtyfour clusters can be linked by way of a Banyan network. Messages are routed through the switch (606) in the form of packets which include a command field, a sequence number, a destination address, a source address, a data field

(which can include subcommands), and an error correction code. Error correction is performed at the processing elements. If a packet is routed to a non-present or non-functional processor, the switch (606) reverses the source and destination field and returns the packet to the sender with an error flag. If the packet is misrouted to a functional processing element (604), the processing element (604) corrects the error and retransmits the packet through the switch (606) over a different path. In one embodiment, each processing element can be provided with a hardware accelerator for database functions. In this embodiment, the multiprocessor of the present invention can be employed as a coprocessor to a 370 host and used to perform database functions.

EP 0 439 693 A3

FIG. 6





European Patent  
Office

# EUROPEAN SEARCH REPORT

Application Number

EP 90 12 0874

Page 1

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. Cl.5)
X	1982 INT'L CONFERENCE ON PARALLEL PROCESSING, Aug 24-27, 1982, K. E. BATCHER ET AL: 'Fault-Tolerant Interconnection Network using Error Correcting Codes', pages 123-125 * page 123, column 1, line 20 - page 125, column 1, line 30 *	1	G06F11/00 G06F11/10
Y		2,3	
A		5,7,17,18	
Y	1986 INT'L CONFERENCE ON PARALLEL PROCESSING Aug 19-22, 1986, K. HWANG ET AL: 'A Fault-Tolerant Interconnection Network Supporting the Fetch-and Add Primitive', pages 327-334 * abstract; figures 2,4,7 * * page 327, column 2, line 7 - page 328, column 2 * * page 331, column 1, paragraph 3 - page 332, column 2 *	2,3	
A		1,5,7,8,10,11,18	TECHNICAL FIELDS SEARCHED (Int. Cl.5)
P,A	INTERNATIONAL JOURNAL OF ELECTRONICS vol. 68, no. 6, June 1990, LONDON, GB pages 901 - 913; K. Y. SRINISVAN, A. K. SOOD: 'Analysis and Design of a Fault-Tolerant Tree Architecture' * page 902, paragraph 2 - page 903, paragraph 1 * * page 910, paragraph 2 - page 911, line 40 *	1-5,18	G06F H04L
A	FTCS-1, The 11th Annual Int'l Symposium on Fault Tolerant Computing, Portland, USA, June 24-26, D. K. PRADHAN, S. M. REDDY: 'A Fault-tolerant Communication Architecture for Distributed * page 214, column 1, line 1 - page 215, column 1, line 22 * * page 216, column 2, line 12 - page 217, column 2; figures 1,4 * Systems', pages 214-220	1,3,5,13,14,18	
The present search report has been drawn up for all claims			
Place of search THE HAGUE		Date of completion of the search 14 APRIL 1992	Examiner JOHANSSON U.C.
<b>CATEGORY OF CITED DOCUMENTS</b> X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons A : member of the same patent family, corresponding document			

EPO FORM 150 (01.87) (P0001)



European Patent  
Office

# EUROPEAN SEARCH REPORT

Application Number

EP 90 12 0874  
Page 2

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int. CL.5)
A	IEEE INFOCOM '88, /th Annual Joint Conference of the IEEE Computer and Communication Societies, New Orleans, USA, March 27-31, 1988, K. V. LE, C. S. RAGHAVENDRA: 'Fault Tolerant Routing in a Class of Double Loop Networks', pages 264-273 * page 266, column 2 - page 267, column 2 * * page 268, column 2, paragraph 3 - page 270, column 1 *  -----	1,3,4,18	
			TECHNICAL FIELDS SEARCHED (Int. CL.5)
The present search report has been drawn up for all claims			
Place of search THE HAGUE		Date of completion of the search 14 APRIL 1992	Examiner JOHANSSON U.C.
<b>CATEGORY OF CITED DOCUMENTS</b> X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document  T : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons  A : member of the same patent family, corresponding document			

EPO FORM 1503 (1.1.92) (P.0001)